State Complexity and the Monoid of Transformations of a Finite Set

Bryan Krawetz
School of Computer Science
bakrawet@uwaterloo.ca

John Lawrence
Department of Pure Mathematics
jwlawren@math.uwaterloo.ca

Jeffrey Shallit
School of Computer Science
shallit@uwaterloo.ca

University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

Abstract

In this paper we consider the state complexity of an operation on formal languages, root(L). This naturally entails the discussion of the monoid of transformations of a finite set. We obtain good upper and lower bounds on the state complexity of root(L) over alphabets of all sizes.

1   Introduction

A *deterministic finite automaton*, or *DFA*, is a 5-tuple $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$, where Q is a finite non-empty set of states, Σ is the finite input alphabet, $\delta : Q \times \Sigma \to Q$ is the transition function, $q_0 \in Q$ is the initial state, and $F \subseteq Q$ is the set of final states. We assume that δ is defined on all elements of its domain. The domain of δ can be extended in the obvious way to $Q \times \Sigma^*$, where Σ^* is the free monoid over the alphabet Σ. For a DFA $\mathcal{A}$, the set $L(\mathcal{A}) = \{w \in \Sigma^* : \delta(q_0, w) \in F\}$ is said to be the language recognized by $\mathcal{A}$.

The *state complexity* of a regular language $L \subseteq \Sigma^*$, denoted sc(L), is defined as the size (the number of states) of the smallest DFA recognizing L. The state complexity of various operations on regular languages, such as union, concatenation, and Kleene closure, has been studied extensively; see, for example, [14, 15].

In this paper we examine a less familiar operation, namely root(L), which is given by

$$\text{root}(L) = \{w \in \Sigma^* : \exists n \geq 1 \text{ such that } w^n \in L\}.$$

Note that this operation is not the same as the ROOT(L) operation studied by Horváth, Leupold, and Lischke [7]. The study of the root(L) operation requires us to examine the connections between finite automata and algebra.

For a finite set Q, a function $f : Q \to Q$ is called a *transformation*. The set of all transformations of Q is denoted Q^Q. For transformations $f, g \in Q^Q$, their composition is written fg, and is given by $(fg)(q) = g(f(q))$, for all $q \in Q$. Together, the set Q^Q and the composition operator form a monoid.

Transformations and their monoids have been studied in some detail by Dénes (whose work is summarized in [3]), and Salomaa [11, 12]. Dénes investigates several algebraic and combinatorial properties of transformations, while much of Salomaa's work is concerned with subsets that generate the full monoid of transformations.

Let L be a language and $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$ a DFA such that $L = L(\mathcal{A})$. For $w \in \Sigma^*$, define $\delta_w(q) = \delta(q, w)$, for all $q \in Q$. Then δ_w is a transformation of Q. If we denote the empty word by ϵ, then δ_ϵ is the identity transformation.

Theorem 1.1. *For a language L and a DFA $\mathcal{A} = (Q, \Sigma, \delta, q_0, F)$ with $L = L(\mathcal{A})$, define the DFA $\mathcal{A}' = (Q^Q, \Sigma, \delta', q_0', F')$ where $q_0' = \delta_\epsilon$, $F' = \{f : \exists n \geq 1 \text{ such that } f^n(q_0) \in F\}$, and δ' is given by*

$$\delta'(f, a) = f\delta_a, \text{ for all } f \in Q^Q \text{ and } a \in \Sigma.$$

Then $\mathrm{root}(L) = L(\mathcal{A}')$.

Proof. An easy induction on $|w|$, $w \in \Sigma^*$, proves that $\delta'(q_0', w) = \delta_w$. Then

$$\begin{aligned}
x \in \mathrm{root}(L) &\Leftrightarrow \exists n \geq 1 \ : \ x^n \in L \\
&\Leftrightarrow \delta_x(q_0) \in F' \\
&\Leftrightarrow \delta'(q_0', x) \in F'.
\end{aligned}$$

$\square$

In addition to giving us a construction for a DFA recognizing $\mathrm{root}(L)$, the above result shows that this operation preserves regularity, that is, if L is regular, then $\mathrm{root}(L)$ is regular. Zhang [16] used a similar technique to characterize regularity-preserving operations. To recognize the image of a language under an operation, Zhang constructs a new automaton with states based on Boolean matrices. These matrices represent the transformations of states in the original automaton.

The result of Theorem 1.1 also allows us to give our first bound on the state complexity of $\mathrm{root}(L)$.

Corollary 1.2. *For regular language L, if $\mathrm{sc}(L) = n$ then $\mathrm{sc}(\mathrm{root}(L)) \leq n^n$.*

Proof. This is immediate from the construction given in Theorem 1.1. $\square$

In the remainder of this paper we improve on this upper bound, and give a non-trivial lower bound for the worst-case blow-up of the state complexity of $\mathrm{root}(L)$, for alphabets of all sizes. These upper and lower bounds demonstrate that a simple, intuitive operation that preserves regularity can increase the state complexity of a language from n to nearly n^n, even over binary alphabets.

Our main results are given in Corollary 3.21, Corollary 3.24, and Theorem 3.27.

2 Unary languages

In the case of unary regular languages, it turns out that the state complexity of the root of a language is bounded by the state complexity of the original language.

Proposition 2.1. *If L is a unary regular language, then $\mathrm{sc}(\mathrm{root}(L)) \leq \mathrm{sc}(L)$. This bound is tight.*

The idea of the following proof is that given a particular DFA recognizing L, we can modify it by adding states to the set of final states. The resulting DFA will recognize the language $\mathrm{root}(L)$.

Proof. Let $\Sigma = \{a\}$ be the alphabet of L. Since L is regular and unary, there exists a DFA $\mathcal{A}$ recognizing L, such that $\mathcal{A} = (\{q_0, \ldots, q_{n-1}\}, \{a\}, \delta, q_0, F)$, where

$$\delta(q_i, a) = q_{i+1}, \text{ for all } 0 \le i < n - 1,$$

and

$$\delta(q_{n-1}, a) = q_j, \text{ for some } 0 \le j \le n - 1.$$

We call the states $q_0, \ldots, q_{j-1}$ the *tail*, and the states $q_j, \ldots, q_{n-1}$ the *loop*.

Notice that $\text{root}(L) = \{a^s \in \Sigma^* : s \mid t, a^t \in L\}$. For all strings $a^t \in L$, we have some $k \ge 0$ and some $b \le n - 1$ such that $t = kl + b$, where $l = n - j$ is the number of states in the loop. Let $s = lm + c$ for some $m \ge 0$ and some $0 \le c < l$. Then

$$
\begin{aligned}
s \mid t \;\Leftrightarrow\; & \exists r \;:\; lk + b = r(lm + c) \\
\Leftrightarrow\; & \exists r \;:\; lk - rlm = rc - b \\
\Leftrightarrow\; & \exists r \;:\; \gcd(l, -lm) \mid rc - b && \text{(by Theorem 4.3.1 of [1])} \\
\Leftrightarrow\; & \exists r \;:\; l \mid rc - b \\
\Leftrightarrow\; & \exists r, v \;:\; rc - b = lv \\
\Leftrightarrow\; & \exists r, v \;:\; rc - lv = b \\
\Leftrightarrow\; & \gcd(l, c) \mid b. && \text{(by Theorem 4.3.1 of [1])}
\end{aligned}
$$

It follows that the set of divisors of the numbers of the form $kl + b$, $k \ge 0$, $b \le n - 1$ is as follows:

$$\{lm + c \in \mathbb{Z} : m \ge 0, \gcd(l, c) \mid b\}.$$

These divisors can be recognized by changing the corresponding states into final states. Therefore, $\text{sc}(\text{root}(L)) \le \text{sc}(L)$.

To show that this bound is tight, for $n \ge 2$ consider the language $L_n = \{a^{n-2}\}$. Under the Myhill-Nerode equivalence relation [6], no two strings in the set $\{\epsilon, a, a^2, \ldots, a^{n-1}\}$ are equivalent. All other strings in Σ^* are equivalent to a^{n-1}. This gives $\text{sc}(L_n) = n$. Furthermore, since a^{n-2} is the longest word in $\text{root}(L_n)$, $\delta(q_0, a^{n-2})$ cannot be a state in the loop. It follows that we require exactly $n - 1$ states in the tail plus a single, non-final state in the loop. Hence $\text{sc}(\text{root}(L_n)) = n$. Therefore the bound is tight. $\square$

3   Languages on larger alphabets

For a regular language $L \subseteq \Sigma^*$, if $\mathcal{A}$ is the minimal DFA such that $L = L(\mathcal{A})$, then as we saw in Section 1, based on the set of all transformations of the states of $\mathcal{A}$, we can construct an automaton $\mathcal{A}'$ as in Theorem 1.1, to recognize $\text{root}(L)$. Though this new DFA, $\mathcal{A}'$, has all transformations of Q as its states, it is easy to see that the only reachable states are those that are a composition of the transformations $\delta_{a_1}, \ldots, \delta_{a_m}$, where $a_1, \ldots, a_m \in \Sigma$. This set of elements, $\delta_{a_1}, \ldots, \delta_{a_m}$, and all of their compositions form the transformation monoid of $\mathcal{A}$. We use this fact to improve on the upper bound of the state complexity of $\text{root}(L)$.

Corollary 3.1. *For a regular language L, let $\mathcal{A}$ be the smallest DFA recognizing L. Then if M is the transformation monoid of $\mathcal{A}$, we have that $\mathrm{sc}(\mathrm{root}(L)) \leq |M|$.*

Proof. In Theorem 1.1, the only reachable states in the construction of $\mathcal{A}'$ are those that belong to the transformation monoid of $\mathcal{A}$. $\qquad\square$

Define $Z_n = \{1, 2, \ldots, n\}$. Now define $T_n = Z_n^{Z_n}$, the set of transformations of Z_n, and $S_n \subseteq T_n$ as the set of permutations of Z_n. For $\gamma \in T_n$ we write

$$\gamma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \gamma(1) & \gamma(2) & \cdots & \gamma(n) \end{pmatrix}.$$

Definition 3.2. If $M \subseteq T_n$ is the set of all compositions of the transformations $f_1, \ldots, f_m \in T_n$, then we say that $\{f_1, \ldots, f_m\}$ *generates* M.

Definition 3.3. For $\gamma \in T_n$, define the *image* of γ by $\mathrm{img}(\gamma) = \{y \in Z_n : y = \gamma(z) \text{ for some } z \in Z_n\}$.

Definition 3.4. For $\gamma \in T_n$, define the *rank* of γ as the number of distinct elements in the image of γ, and denote it by $\mathrm{rank}(\gamma)$.

The relationship between the state complexity of a language and the transformation monoid naturally leads to the question of how large a submonoid of T_n can be generated by m elements, where m is a positive integer. In connection with the study of Landau's function (for a survey see [9, 10]), Szalay [13] showed that, for $m = 1$, the largest submonoid of T_n has size

$$\exp\left\{ \sqrt{n\left(\log n + \log\log n - 1 + \frac{\log\log n - 2 + o(1)}{\log n}\right)} \right\}.$$

In the case where $m \geq 3$, the results are well known.

Lemma 3.5. *For $n \geq 3$, suppose $H \subseteq T_n$ such that H generates T_n. Then $|H| \geq 3$. Furthermore, $|H| = 3$ if and only if H can be written as $H = \{\alpha, \beta, \gamma\}$, where $\{\alpha, \beta\}$ generates S_n and $\mathrm{rank}(\gamma) = n - 1$.*

For a proof of this lemma, see Dénes [2]. This gives us the result that the largest submonoid generated by three elements has the full size n^n.

Contrary to the case for $m = 1$ and $m \geq 3$, it seems that only recently there has been any interest in determining the largest submonoid on two generators. Significant progress has been made in this area by Holzer and König [4, 5], and, independently, by Krawetz, Lawrence, and Shallit [8]. The results of Holzer and König are summarized here.

For coprime integers $k, l \geq 2$, where $k + l = n$, let $\alpha = (1\ 2\ \cdots\ k)(k+1\ \ k+2\ \ \cdots\ n)$ be a permutation of Z_n composed of two cycles, one of length k, the other of length l. Define $U_{k,l}$ to be the set of all transformations $\gamma \in T_n$ where exactly one of the following is true:

    1. $\gamma = \alpha^m$ for some positive integer m;

2. For some $i \in \{1, \cdots, k\}$ and some $j \in \{k+1, \cdots, n\}$ we have that $\gamma(i) = \gamma(k)$ and for some $m \in \{k+1, \cdots, n\}$ we have that $m \notin \text{img}(\gamma)$.

Let $\pi_1 = (1 \ 2 \ \cdots \ k)$ be an element of S_{n-1}, and let $\pi_2 \in S_{n-1}$ be a permutation such that π_1 and π_2 generate S_{n-1}. Now define $\beta \in T_n$ by

$$\beta = \begin{pmatrix} 1 & 2 & \cdots & n-1 & n \\ \pi_2(1) & \pi_2(2) & \cdots & \pi_2(n-1) & \pi_2(1) \end{pmatrix}.$$

Lemma 3.6 (Holzer and König). *The set $U_{k,l}$ is a submonoid of T_n and is generated by $\{\alpha, \beta\}$.*

It is worth noting that in their definition of $U_{k,l}$, Holzer and König allow $k = 1$ and $l = 1$. They show implicitly, however, that the size of the monoid in these degenerate cases is too small to be of any consequence here.

Theorem 3.7 (Holzer and König). *For $n \geq 7$, there exist coprime integers k,l such that $n = k+l$ and*

$$|U_{k,l}| \geq n^n \left(1 - \sqrt{2} \left(\frac{2}{e} \right)^{\frac{n}{2}} e^{\frac{1}{12}} - \sqrt{8} \frac{1}{\sqrt{n}} e^{\frac{1}{12}} \right).$$

In addition to a lower bound on the size of the largest two-generated monoid, Theorem 3.7 gives us the existence of a sequence of two-generated monoids whose size approaches n^n as n tends toward infinity. Similar results were obtained independently by Krawetz, Lawrence, and Shallit [8].

More recently, Holzer and König [5] proved the following result regarding the maximality of monoids of the form $U_{k,l}$.

Theorem 3.8 (Holzer and König). *For all prime numbers $n \geq 7$, there exist coprime integers $k,l \geq 2$ such that $k + l = n$ and $U_{k,l}$ is the largest two-generated submonoid of T_n.*

They also stated the following conjecture.

Conjecture 3.9 (Holzer and König). *For any $n \geq 7$, there exist coprime integers $k,l \geq 2$ such that $k + l = n$ and $U_{k,l}$ is the largest two-generated submonoid of T_n.*

Since the connection between the state complexity of $\text{root}(L)$ and the transformation monoid of L has been established in Corollary 3.1, we can take advantage of the results of Theorem 3.7 and Theorem 3.8 if we can construct a language based on a monoid. By associating an alphabet with the generators of a monoid, we can define a transition function for a DFA. The definition of the DFA is then completed by choosing a start state and a set of final states. This construction is given more formally below.

Let n, m be integers with $n, m \geq 1$. For a set of transformations $X = \{\alpha_1, \ldots, \alpha_m\}$, let $M \subseteq T_n$ denote the monoid generated by X. Then a *DFA based on X* is a DFA $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, F)$, where $|\Sigma| \geq m$, $z_0 \in Z_n$, $F \subseteq Z_n$, and δ is given by

$$\delta_a = \Psi(a), \text{ for all } a \in \Sigma,$$

for some map $\Psi : \Sigma \to X \cup \{\delta_\epsilon\}$ that is surjective on X.

Proposition 3.10. *Let $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, F)$ be a DFA. Then M is the transformation monoid of $\mathcal{M}$ if and only if $\mathcal{M}$ is based X, for some $X \subseteq T_n$ that generates the monoid M.*

Proof. For a DFA $\mathcal{M}$ based on X, the fact that M is the transformation monoid of $\mathcal{M}$ is immediate from the construction. For any DFA $\mathcal{M}$ that has M as its transformation monoid, we have that the set $\{\delta_a \in T_n : a \in \Sigma\}$ generates M. Then we can simply take Ψ given by $\Psi(a) = \delta_a$, for all $a \in \Sigma$. $\qquad\square$

In particular, let $\mathcal{A}_{\Psi,X} = (Z_n, \Sigma, \delta, z_0, F)$ denote the DFA based on X when $z_0 = 1$, $F = \{1\}$, and Ψ is bijective on an m-element subset of Σ, with all other elements of Σ mapped to δ_ϵ. If Ψ_1 and Ψ_2 are maps over the same domain, then $\mathcal{A}_{\Psi_1,X}$ is isomorphic to $\mathcal{A}_{\Psi_2,X}$, up to a renaming of the states and alphabet symbols. For this reason, we will often denote this DFA simply by $\mathcal{A}_{\Sigma,X}$.

Example 3.11. Let $Y = \{\alpha, \beta\}$, where

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{pmatrix}, \text{ and } \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 1 & 2 \end{pmatrix}.$$

Define Φ by $\Phi(a) = \alpha$ and $\Phi(b) = \beta$. Then Figure 3.12 depicts the DFA $\mathcal{A}_{\Phi,Y}$.
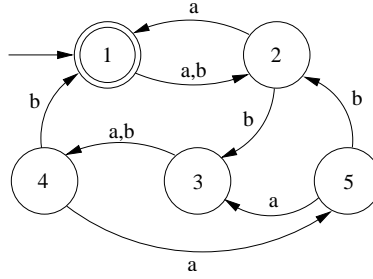

Figure 3.12: The automaton $\mathcal{A}_{\Phi,Y}$.

For $n \geq 5$, define $X_n \subseteq T_n$ to be a subset of transformations, and let M_n denote the monoid generated by X_n. In particular, define $X_{k,l} = \{\alpha, \beta\}$, where α and β are as in Lemma 3.6. Then $X_{k,l}$ generates $U_{k,l}$. We now state our main result concerning the state complexity of $\text{root}(L(\mathcal{A}_{\Sigma,X_n}))$.

Theorem 3.13. *If $U_{k,l} \subseteq M_n$, for some coprime integers $k \geq 2$ and $l \geq 3$, with $k + l = n$, then the minimal DFA recognizing $\text{root}(L(\mathcal{A}_{\Sigma,X_n}))$ has $|M_n| - \binom{n}{2}$ states.*

Before we are ready to prove this theorem, we must state a few more definitions and lemmas.

Definition 3.14. Let $\rho \in T_n$. For any i, j, k, if $\rho(i) = k = \rho(j)$ implies that $i = j$, then we say that k is *unique*.

Definition 3.15. Let $\rho \in T_n$ have rank 2, with $\text{img}(\rho) = \{i, j\}$. Then by the *complement* of ρ, we mean the transformation $\overline{\rho} \in T_n$, where

$$\overline{\rho}(k) = \begin{cases} i, & \text{if } \rho(k) = j; \\ j, & \text{if } \rho(k) = i. \end{cases}$$

6

For example, if $\rho = \begin{pmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ 3 & 3 & 2 & \cdots & 2 & 2 \end{pmatrix}$, then $\overline{\rho} = \begin{pmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ 2 & 2 & 3 & \cdots & 3 & 3 \end{pmatrix}$.

It is easy to see that, in general, ρ and $\overline{\rho}$ have the same rank, and $\overline{\overline{\rho}} = \rho$.

For an automaton $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, F)$, define the DFA $\mathcal{M}^* = (M_n, \Sigma, \delta', \delta_\epsilon, F')$, where δ_ϵ is the identity element of T_n, $\delta'(\eta, a) = \eta \delta_a$ for all $\eta \in M_n$, for all $a \in \Sigma$, and $F' = \{\eta \in T_n : \eta(z_0) = z_0\}$. Then $L(\mathcal{M}^*) = \mathrm{root}(L(\mathcal{M}))$.

For $\eta, \theta \in M_n$, with $\eta \neq \theta$, note that η and θ are equivalent states if and only if for all $\rho \in \Sigma^*$ we have that
$$\delta'(\eta, \rho) \in F' \Leftrightarrow \delta'(\theta, \rho) \in F'.$$
However, since $\delta'(\eta, \rho) = \eta \delta_\rho$, this is equivalent to saying that η and θ are equivalent states in M if and only if for all $\rho \in M_n$, we have
$$\eta\rho \in F' \Leftrightarrow \theta\rho \in F'.$$

Lemma 3.16. *Let $Y \subseteq T_n$ generate M_n, and let $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, F)$ be an automaton based on Y such that $z_0 \in F$. Let $\eta, \theta \in M_n$, with $\eta \neq \theta$ and $\mathrm{rank}(\eta) = 2$. If $\eta(z_0)$ is unique in the image of η, and $\overline{\eta} = \theta$, then η and θ are equivalent states in $\mathcal{M}^*$.*

Proof. We have
$$\eta = \begin{pmatrix} 1 & 2 & \cdots & z_0 - 1 & z_0 & z_0 + 1 & \cdots & n \\ j & j & \cdots & j & i & j & \cdots & j \end{pmatrix},$$
and
$$\theta = \begin{pmatrix} 1 & 2 & \cdots & z_0 - 1 & z_0 & z_0 + 1 & \cdots & n \\ i & i & \cdots & i & j & i & \cdots & i \end{pmatrix},$$
for some $i \neq j$.

If $\eta(z_0) \in F$, then $\eta \in F'$. If $\theta(z_0) \in F$, then $\theta \in F'$. Otherwise, $\theta(z_0) \in Z_n \backslash \{z_0\}$. Since $\theta(z) = \eta(z_0) \in F$ for all $z \in Z_n \backslash \{z_0\}$, we have that $\theta^2(z_0) \in F$, and hence $\theta \in F'$. Similarly, $\theta(z_0) \in F$ implies that $\eta, \theta \in F'$. Furthermore, if $\mathrm{img}(\eta) \cap F = \emptyset$, then $\mathrm{img}(\eta^n) \cap F = \emptyset$, for all n, and hence $\eta^n(z_0) \notin F$ so that $\eta \notin F'$. Since $\mathrm{img}(\eta) = \mathrm{img}(\theta)$, this gives $\theta \notin F'$. Therefore $\eta \in F'$ if and only if $\theta \in F'$.

Let $\rho \in M_n$. Since η and θ have rank 2, we must have that $\eta\rho$ and $\theta\rho$ have rank ≤ 2. If $\eta\rho$ has rank 2, then $\rho(i) \neq \rho(j)$, so that $\overline{\eta\rho} = \theta\rho$. Hence $\eta\rho \in F'$ if and only if $\theta\rho \in F'$. The argument is the same for the case where $\theta\rho$ has rank 2. Now, if $\eta\rho$ has rank 1, then we must have $\rho = \rho'\sigma\rho''$, where $\sigma(s) = \sigma(t)$ for some s and t, and where ρ' is a permutation such that either $\eta\rho'(z_0) = s$ and $\eta\rho'(z) = t$, for all $z \neq z_0$, or $\eta\rho'(z_0) = t$ and $\eta\rho'(z) = s$, for all $z \neq z_0$. Without loss of generality, assume the former. Then clearly $\theta\rho'(z_0) = t$ and $\theta\rho'(z) = s$, for all $z \neq z_0$. It follows that $\eta\rho'\sigma = \theta\rho'\sigma$, so that $\eta\rho = \theta\rho$.

Therefore η and θ are equivalent states. $\qquad\square$

Lemma 3.17. *Let $Y \subseteq T_n$ generate M_n, and let $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, F)$ be an automaton based on Y, such that $z_0 \notin F$. Let $\eta, \theta \in M_n$, with $\eta(z_0)$ unique in the image of η, $\mathrm{rank}(\eta) = 2$, and $\mathrm{img}(\eta) = \mathrm{img}(\theta)$. If $\theta(z_0) = \eta(z_0)$, and if $\theta(z) = \eta(z_0)$ implies that $z \in F$, then η and θ are equivalent states in $\mathcal{M}^*$.*

Proof. If $\eta(z_0) \in F$, then $\eta \in F'$. Since $\theta(z_0) = \eta(z_0)$, it follows that $\theta \in F'$. Now suppose that $\eta(z_0) \notin F$. If $\eta \in F'$, then since $\mathrm{rank}(\eta) = 2$, we must have that $\eta^2(z_0) \in F$. Now $\theta(z_0) \notin F$, so $\theta(\theta(z_0)) \neq \eta(z_0)$. But since $\mathrm{rank}(\theta) = 2$, and $\mathrm{img}(\theta) = \mathrm{img}(\eta)$, it follows that $\theta^2(z_0) = \eta^2(z_0)$. Hence $\theta \in F'$. If $\eta \notin F'$, then $\eta(z_0) = z_0$ or $\mathrm{img}(\eta) \cap F = \emptyset$. In either case, this implies that $\theta \notin F'$. Therefore $\eta \in F'$ if and only if $\theta \in F'$.

Let $\rho \in M_n$. Then, following an argument similar to the one used in the proof of Lemma 3.16, we have that $\eta\rho \in F'$ if and only if $\theta\rho \in F'$. Therefore η and θ are equivalent states. $\square$

Now that we have a characterization of equivalent states in the general case for $\mathcal{M}^*$, we turn our attention toward the specific case, for $\mathcal{A}^*_{\Sigma, X_n}$.

Lemma 3.18. *Let $\eta, \theta \in M_n$, with $\eta \neq \theta$. If $\mathrm{rank}(\eta) = 1$, then η and θ are not equivalent states in $\mathcal{A}^*_{\Sigma, X_n}$.*

Proof. Since η has rank 1, we have that $\mathrm{img}(\eta) = \{z_1\}$ for some z_1. If $\eta(1) \neq \theta(1)$, then take $\rho \in U_{k,l}$ such that $\rho(z_1) = 2$, and $\rho(z) = 1$, for all $z \neq z_1$. Then $\mathrm{img}(\eta\rho) = \{2\}$, so that $\eta\rho \notin F'$. But $\theta\rho(1) = 1$, so that $\theta\rho \in F'$. Hence η and θ are not equivalent. If $\eta(1) = \theta(1)$, then $\mathrm{rank}(\theta) \neq 1$ so that for some $z_2 \neq 1$ we have $\theta(z_2) \neq z_1$. Take $\rho \in U_{k,l}$ such that $\rho(\theta(z_2)) = 1$, and $\rho(z) = z_2$, for all $z \neq \theta(z_2)$. Then $\mathrm{img}(\eta\rho) = \{z_2\}$, so that $\eta\rho \notin F'$. But $(\theta\rho)^2(1) = 1$, so that $\theta\rho \in F'$. Hence η and θ are not equivalent. $\square$

Lemma 3.19. *For $\eta, \theta \in M_n$, with $\eta \neq \theta$, let η have rank 2. Then η and θ are equivalent states in $\mathcal{A}^*_{\Sigma, X_n}$ if and only if $\eta(1)$ is unique in the image of η, and $\overline{\eta} = \theta$.*

Proof. Let $\mathrm{img}(\eta) = \{i, j\}$ for some i, j. Without loss of generality, assume that $\eta(1) = i$.

Since $1 \in F$, Lemma 3.16 applies and gives the result in the forward direction. For the other direction, we have two cases.

Case 1. i is not unique in the image of η.

    Case 1.a. $i = \eta(1) \neq \theta(1)$.

        Since i is not unique, $\eta(z) = i$, for some $z \neq 1$. Choose $\rho \in U_{k,l}$ such that $\rho(i) = z$, and $\rho(\theta(1)) = 1$. Then $\eta\rho(1) = \eta\rho(z) = z$ so that $(\eta\rho)^n(1) = z$, for all $n \geq 0$. This gives us $\eta\rho \notin F'$. But $\theta\rho(1) = 1$, so that $\theta\rho \in F'$.

    Case 1.b. $i = \eta(1) = \theta(1)$.

        Case 1.b.i. $\mathrm{img}(\eta) = \mathrm{img}(\theta)$.

            Since $\mathrm{img}(\eta) = \mathrm{img}(\theta)$ and $\eta \neq \theta$, then for some $z \neq 1$, we must have either $i = \eta(z) \neq \theta(z) = j$, or $j = \eta(z) \neq \theta(z) = i$. Without loss of generality, assume the former. Then choose $\rho \in U_{k,l}$ such

8

that $\rho(i) = z$, and $\rho(j) = 1$. Then $(\eta\rho)^n(1) = z$, for all $n \geq 0$, and $(\theta\rho)^2(1) = 1$. This gives us $\eta\rho \notin F'$ and $\theta\rho \in F'$.

Case 1.b.ii. $\text{img}(\eta) \neq \text{img}(\theta)$.

If $\text{rank}(\theta) = 1$ then by Lemma 3.18, η and θ are not equivalent. Otherwise there exists some $z_1 \in Z_n$ such that $\theta(z_1) \notin \text{img}(\eta)$. Take $\rho \in U_{k,l}$ such that $\rho(\theta(z_1)) = 1$, and $\rho(z) = 2$, for all $z \neq \theta(z_1)$. Then $\eta\rho \neq \theta\rho$ and $\text{rank}(\eta\rho) = 1$, and so by Lemma 3.18, $\eta\rho$ and $\theta\rho$ are not equivalent. Hence η and θ are not equivalent.

Case 2. i is unique in the image of η, and $\overline{\eta} \neq \theta$.

If $\text{rank}(\theta) = 1$ then by Lemma 3.18, η and θ are not equivalent. If $\text{img}(\eta) = \text{img}(\theta)$, then since $\overline{\eta} \neq \theta$, we have that $\theta(1)$ is not unique. So we can reverse the roles of η and θ and apply case 1 to get the desired result. Assume then, that $\text{img}(\eta) \neq \text{img}(\theta)$, and $\text{rank}(\theta) \geq 2$. Then the fact that η and θ are not equivalent follows just as in case 1.b.ii.

Therefore, η, and θ are equivalent states if and only if $\eta(1)$ is unique in the image of η, and $\overline{\eta} = \theta$. $\qquad\square$

Lemma 3.20. *Let $\eta, \theta \in M_n$, with $\eta \neq \theta$, if η, θ have rank ≥ 3, then η and θ are not equivalent states in $\mathcal{A}_{\Sigma, X_n}^*$.*

Proof. Since $\eta \neq \theta$, there exists some $z_1 \in Z_n$ such that $\eta(z_1) \neq \theta(z_1)$. Let $z_2 = \eta(z_1)$. Take $\rho \in U_{k,l}$ such that $\rho(z_2) = 1$, and $\rho(z) = 2$, for all $z \neq z_2$. Since $\text{rank}(\eta) \geq 3$, we have $rank(\eta\rho) = 2$. If $\eta\rho(1)$ is not unique, then by Lemma 3.19, $\eta\rho$ and $\theta\rho$ are not equivalent. Hence η and θ are not equivalent. If $\eta\rho(1)$ is unique, then it must be that $z_1 = 1$. Furthermore, since $\text{rank}(\theta) \geq 3$, we cannot have $\theta(z) = z_2$ for all $z \neq 1$, so we cannot have $\theta\rho(z) = 1$ for all $z \neq 1$. Therefore $\theta\rho \neq \overline{\eta\rho}$. Then by Lemma 3.19, $\eta\rho$ and $\theta\rho$ are not equivalent. Hence η and θ are not equivalent. $\qquad\square$

We are now ready to prove Theorem 3.13.

Proof (Theorem 3.13). Lemma 3.18 – 3.20 cover all possible cases for $\eta, \theta \in M_n$, $\eta \neq \theta$. Therefore, two states are equivalent if and only if they satisfy the hypothesis of Lemma 3.16. There are $\binom{n}{2}$ such equivalence classes in $U_{k,l} \subseteq M_n$, each containing exactly 2 elements. All other elements of M_n are in equivalence classes by themselves. It follows that the minimal DFA recognizing $\text{root}(L(\mathcal{A}_{\Sigma, X_n}))$ has $|M_n| - \binom{n}{2}$ states. $\qquad\square$

Now that we have established a close relationship between $\text{sc}(\text{root}(L))$ and the transformation monoid of the the minimal automaton recognizing L, we can take advantage of results concerning the size of the largest monoids to give bounds on the worst-case blow-up of the state complexity of $\text{root}(L)$. The following corollary gives a lower bound for alphabets of size two. It also proves the existence of a sequence of regular binary languages with state complexity n whose root has a state complexity that approaches n^n as n increases without bound.

We now state our first main result.

Corollary 3.21. *For $n \geq 7$, there exists a regular language L over an alphabet of size 2, with $\mathrm{sc}(L) \leq n$, such that*

$$\mathrm{sc}(\mathrm{root}(L)) \geq n^n \left(1 - \sqrt{2} \left(\frac{2}{e} \right)^{\frac{n}{2}} e^{\frac{1}{12}} - \sqrt{8} \frac{1}{\sqrt{n}} e^{\frac{1}{12}} \right) - \binom{n}{2}.$$

Proof. The result follows from a combination of Theorem 3.7 and Theorem 3.13. $\qquad \square$

Our results from Theorem 3.13 do not apply when $l = 2$. Unfortunately, Theorem 3.8 does not exclude this possibility. To guarantee that this fact is of no consequence, we must show that not only is the monoid $U_{n-2,2}$ never the largest, but that it is at least $\binom{n}{2}$ smaller than the largest monoid. The following lemma deals with this.

Lemma 3.22. *For $n \geq 7$, we have that*

$$|U_{2,n-2}| - |U_{n-2,2}| \geq \binom{n}{2}.$$

Due to space constraints, the proof of this lemma has been relegated to the appendix.

The choice of start and final states in the construction of the DFA $\mathcal{A}_{\Sigma, X_n}$ is the best possible. The following theorem will show that for any other DFA with the same transition function, a different assignment of start and final states will not increase the state complexity of the language it recognizes.

Theorem 3.23. *Let $Y \subseteq T_n$ generate M_n, and let $\mathcal{M} = (Z_n, \Sigma, \delta, z_0, G)$ be an automaton based on Y. Then $\mathrm{sc}(\mathrm{root}(L(\mathcal{M}))) \leq \mathrm{sc}(\mathrm{root}(L(\mathcal{A}_{\Sigma, X_n})))$.*

Proof. If $z_0 \in G$, then Lemma 3.16 applies. It follows that there are at least $\binom{n}{2}$ pairs of equivalent states in M^*. If $z_0 \notin G$, then Lemma 3.17 applies, and again we have at least $\binom{n}{2}$ pairs of equivalent states in M^*. In either case, this gives

$$\mathrm{sc}(\mathrm{root}(L(\mathcal{M}))) \leq |M_n| - \binom{n}{2} \leq \mathrm{sc}(\mathrm{root}(L(\mathcal{A}_{\Sigma, X_n}))).$$

$\qquad \square$

We now state our second main result.

Corollary 3.24. *For prime numbers $n \geq 7$, there exist positive, coprime integers $k \geq 2$, $l \geq 3$, with $k + l = n$, such that if L is a language over an alphabet of size 2, with $\mathrm{sc}(L) \leq n$, then $\mathrm{sc}(\mathrm{root}(L)) \leq |U_{k,l}| - \binom{n}{2}$. Furthermore, this bound is tight.*

Proof. Let U' denote the largest two-generated submonoid of T_n. Then by Theorem 3.8 and Lemma 3.22, we have that $U' = U_{k',l'}$ for some coprime integers $k' \geq 2$, $l' \geq 3$ with $k' + l' = n$.

Let $\mathcal{M}$ be the smallest DFA recognizing L, and let M be the transformation monoid of $\mathcal{M}$. If M is of the form $U_{k,l}$, with $k \geq 2$, $l \geq 3$, then $|U_{k,l}| \leq |U'|$. It follows from Theorem 3.23 that $\mathrm{sc}(\mathrm{root}(L)) \leq |U_{k,l}| - \binom{n}{2} \leq |U'| - \binom{n}{2}$. If M is of the form $U_{k,l}$, with $k = n - 2$, $l = 2$, then by

Corollary 3.1 and Lemma 3.22 we have $\mathrm{sc}(\mathrm{root}(L)) \leq |U_{n-2,2}| \leq |U_{2,n-2}| - \binom{n}{2} \leq |U'| - \binom{n}{2}$.

Let V denote the largest two-generated submonoid of T_n that is not of the form $U_{k,l}$ for some coprime integers $k \geq 2$, $l \geq 3$ with $k + l = n$. Then for all integers $n > 81$, a simple observation of Holzer and König's proof of Theorem 3.8 shows that $|U'| - |V| \geq \binom{n}{2}$. For all integers $7 \leq n \leq 81$, the fact that $|U'| - |V| \geq \binom{n}{2}$ has been verified computationally. It follows from Corollary 3.1 that if M is not of the form $U_{k,l}$, we have

$$\mathrm{sc}(\mathrm{root}(L)) \leq |M| \leq |V| \leq |U_{k,l}| - \binom{n}{2}.$$

The fact that the bound is tight is an immediate consequence of Theorem 3.13. $\qquad \square$

If Conjecture 3.9 is true, then for all $n \geq 7$, where n is not prime, the construction of $\mathcal{A}_{\Sigma, X_{k,l}}$ yields a language that is within $\binom{n}{2}$ of the maximum blow-up. We conjecture that this construction achieves the maximum.

Conjecture 3.25. *For and integer $n \geq 7$, there exist positive, coprime integers $k \geq 2$, $l \geq 3$, with $k + l = n$, such that if L is a language over an alphabet of size 2, with $\mathrm{sc}(L) \leq n$, then $\mathrm{sc}(\mathrm{root}(L)) \leq |U_{k,l}| - \binom{n}{2}$. This bound is tight.*

The results concerning the largest monoid on ≥ 3 generators are definite and much simpler. For this reason, on alphabets of size ≥ 3 we are able to give a much better bound.

Lemma 3.26. *For $n \geq 1$, if $M \subseteq T_n$ is a monoid such that $|M| > n^n - \binom{n}{2}$, then $M = T_n$.*

Proof. For $1 \leq n \leq 3$, the result can easily be verified computationally, so assume that $n \geq 4$.

There are $\binom{n}{2}$ transpositions in T_n. Since $|M| > |T_n| - \binom{n}{2}$, it follows that M contains at least one transposition. There are $(n-1)!$ permutations of Z_n that have one cycle of length n. Since $n \geq 4$, we have that $(n-1)! \geq \binom{n}{2}$. Again, considering the size of M, it follows that M contains at least one permutation that is a full n-cycle. It follows that $S_n \subseteq M$.

Furthermore, there are $\binom{n}{2} \cdot n!$ transformations of Z_n that have rank $n - 1$, so it follows that M contains at least one transformation of rank $n-1$. Then by Lemma 3.5, we have that $M = T_n$. $\qquad \square$

We now state our third main result.

Theorem 3.27. *Let Σ be an alphabet of size $m \geq 3$. For $n \geq 1$, if L is a language over Σ with $\mathrm{sc}(L) \leq n$, then $\mathrm{sc}(\mathrm{root}(L)) \leq n^n - \binom{n}{2}$. Furthermore, this bound is tight.*

Proof. Define M to be the transformation monoid of the smallest DFA recognizing L. If $|M| \leq n^n - \binom{n}{2}$, then certainly $\mathrm{sc}(\mathrm{root}(L)) \leq n^n - \binom{n}{2}$. So suppose that $|M| > n^n - \binom{n}{2}$. Then it follows from Lemma 3.26 that $M = T_n$.

For $1 \leq n \leq 6$, it has been verified computationally that if the transformation monoid of the minimal DFA recognizing L is T_n, then $\mathrm{sc}(\mathrm{root}(L)) = n^n - \binom{n}{2}$. For $n \geq 7$, if the transformation monoid is T_n, then clearly $U_{k,l} \subseteq T_n$ for some suitable k, l so that Theorem 3.13 applies, and hence

$\text{sc}(\text{root}(L)) = n^n - \binom{n}{2}$.

To show that the bound is tight, it suffices to show that for any n there exists a language L over Σ such that the transformation monoid of the minimal DFA recognizing L is T_n. Let X be a set of transformations such that $|X| = \min(n, 3)$ and X generates T_n. For $n \in \{1, 2\}$, the fact that such an X exists is easy to check. For $n \geq 3$, the existence of X follows from Lemma 3.5. Then the language $L(\mathcal{A}_{\Sigma,X})$, gives the desired result. $\square$

References

[1] E. Bach, and J. Shallit. *Algorithmic Number Theory, Volume 1: Efficient Algorithms.* The MIT Press, 1996.

[2] J. Dénes. On transformations, transformation-semigroups and graphs. In P. Erdös and G. Katona, editors, *Theory of Graphs: Proc. Colloq. Graph Theory (Tihany 1966).* pp 65–75. Academic Press, 1968.

[3] J. Dénes. On a generalization of permutations: some properties of transformations. In *Permutations: Actes du Colloque sur Les Permutations, (Paris 1972)*, pp. 117–120. Gauthier-Villars, 1972.

[4] M. Holzer, and B. König. On deterministic finite automata and syntactic monoid size. In M. Ito, and M. Toyama, editors, *Proceedings of DLT 2002*, Vol. 2450 of *Lecture Notes in Computer Science*, pp. 229–240. Springer-Verlag, 2003.

[5] M. Holzer, and B. König. On deterministic finite automata and syntactic monoid size, continued. In Z. Ésik, and Z. Fülöp, editors, *Proceedings of DLT 2003*, Vol. 2710 of *Lecture Notes in Computer Science*, pp. 349–360. Springer-Verlag, 2003.

[6] J. E. Hopcroft, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, 1979.

[7] S. Horváth, P. Leupold, and G. Lischke. Roots and powers of regular languages. In M. Ito, and M. Toyama, editors, *Proceedings of DLT 2002*, Vol. 2450 of *Lecture Notes in Computer Science*, pp. 220–230. Springer-Verlag, 2002.

[8] B. Krawetz, J. Lawrence, and J. Shallit. State complexity and the monoid of transformations of a finite set. Preprint. Available at http://arxiv.org/math/0306416.

[9] W. Miller. The maximum order of an element of a finite symmetric group. *Amer. Math. Monthly* **94** (1987), 497–506.

[10] J.-L. Nicolas. On Landau's function $g(n)$. In R. L. Graham, and J. Nesetril, editors, *The Mathematics of Paul Erdös*, pp. 228–240. Springer-Verlag, 1997.

[11] A. Salomaa. On basic groups for the set of functions over a finite domain. *Ann. Acad. Scient. Fenn.*, Ser A. I. 338 (1963).

[12] A. Salomaa. Composition sequences for functions over a finite domain. *Theoret. Comp. Sci.* **292** (2003), 263–281.

[13] M. Szalay. On the maximal order in S_n and S_n^*. *Acta Arith.* **37** (1980), 321–331.

[14] S. Yu. State complexity of regular languages. *J. Aut. Lang. and Comb.* **6** (2001), 221–234.

[15] S. Yu, Q. Zhuang, and K. Salomaa. The state complexities of some basic operations on regular languages. *Theoret. Comp. Sci.* **125** (1994), 315–328.

[16] G.-Q. Zhang. Automata, boolean matrices, and ultimate periodicity. *Inform. Comput.* **152** (1999), 138–154.

Appendix: Omitted Proofs

Proof (Lemma 3.22). As stated in [5], for $k + l = n$, we have the following formula

$$|U_{k,l}| = kl + \sum_{i=1}^{n}\left(\binom{n}{i} - \binom{k}{i-l}\right)\left(\left\{{n \atop i}\right\} - \sum_{r=1}^{i}\left\{{k \atop r}\right\}\left\{{l \atop i-r}\right\}\right)i!,$$

where $\left\{{n \atop i}\right\}$ is a Stirling number of the second kind, the number of ways to partition a set of n elements into i non-empty sets. This gives

$$|U_{k,l}| - |U_{l,k}| = \sum_{i=1}^{n}\left(\binom{l}{i-k} - \binom{k}{i-l}\right)\left(\left\{{n \atop i}\right\} - \sum_{r=1}^{i}\left\{{k \atop r}\right\}\left\{{l \atop i-r}\right\}\right)i!. \qquad (*)$$

Since $\left\{{n \atop k}\right\} = 0$ whenever $k > n$ or $k < 1$, for $k = 2$, and $l = n - 2$, we have

$$\sum_{r=1}^{i}\left\{{2 \atop r}\right\}\left\{{n-2 \atop i-r}\right\} = \left\{{n-2 \atop i-2}\right\} + \left\{{n-2 \atop i-1}\right\}.$$

Also, notice that $\binom{n-2}{i-2} - \binom{2}{i-n+2}$ is positive when $2 \geq i \geq n - 1$, and zero otherwise, so that $(*)$ becomes

$$|U_{2,n-2}| - |U_{n-2,2}| \geq \sum_{i=2}^{n-1}\left(\left\{{n \atop i}\right\} - \left\{{n-2 \atop i-2}\right\} - \left\{{n-2 \atop i-1}\right\}\right)i!.$$

And finally, using the identity $\left\{{n \atop k}\right\} = \left\{{n-1 \atop k-1}\right\} + k\left\{{n-1 \atop k}\right\}$, we see that

$$\left\{{n \atop i}\right\} = \left\{{n-2 \atop i-2}\right\} + (2i-1)\left\{{n-2 \atop i-1}\right\} + (i-1)\left\{{n-2 \atop i}\right\},$$

so that we get

$$|U_{2,n-2}| - |U_{n-2,2}| \geq \sum_{i=2}^{n-1}\left((2i-2)\left\{{n-2 \atop i-1}\right\} + (i-1)\left\{{n-2 \atop i}\right\}\right)i! \geq \sum_{i=2}^{n-1}i! \geq \binom{n}{2}.$$

$\square$